

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

Heijne, J., Ungruh, R., & Pera, M. S. (2026). Blurred Lines: Understanding the Fit of Song Lyrics in Music Catalogs That Can Reach Children Through Recommendations. In A. Bellogin, L. Boratto, F. Cena, A. Geninatti Cossatin, T. Huibers, S. Kleanthous, M. Landoni, E. Lex, F. Maridina Malloci, & More Editors (Eds.), *Advances in Bias, Fairness, and Understudied Users in Information Retrieval - 6th International Workshop, BIAS 2025, and 2nd International Workshop, IR4U2 2025, Revised Selected Papers* (pp. 60-75). (Communications in Computer and Information Science; Vol. 2786 CCIS). Springer. https://doi.org/10.1007/978-3-032-12717-4_5

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



Blurred Lines: Understanding the Fit of Song Lyrics in Music Catalogs That Can Reach Children Through Recommendations

Jasper Heijne¹ , Robin Ungruh^(✉) , and Maria Soledad Pera²

Delft University of Technology, Delft, The Netherlands

J.Heijne-1@student.tudelft.nl, {R.Ungruh,M.S.Pera}@tudelft.nl

Abstract. Recommender systems on popular online platforms expose impressionable and easily influenced younger listeners to varied content, making it crucial to reflect on the songs children can encounter due to their interactions with recommender systems. To set a foundation, we analyze the lyrics of a catalog comprised of $\sim 30,000$ songs to gauge their suitability to children. Our multi-perspective exploration reveals a high prevalence of inappropriate lyrics in music commonly heard by children. This highlights the need for further explorations pertaining online platforms and their recommender systems that curate and ultimately present items from catalogs such as the ones we examined, highlighting the potential negative impact of such lyrics on their behavior and personality by promoting harmful language or biases. Informed by our findings, we outline research directions for the information retrieval community to consider when designing, evaluating, and deploying algorithms that serve diverse audiences.

Keywords: Children · Music Recommender Systems · Harmful Content

1 Introduction

Music consumption plays a crucial role in shaping children's¹ cultural identity [57]. From a young age, children link emotions to music, associating it with enjoyment, mood, and social relationships [17, 41]. Musical activities impact their development [43], fostering improved social skills, intelligence, and creativity. However, they can also prompt adverse behaviors and attitudes, e.g., imitating

¹ Whenever we mention children, we refer to individuals aged up to 18, as per UNICEF's definition of a child [91]. In practice, we mean those who independently access online platforms. We recognize crucial developmental and cultural differences among children [44, 92]. Nevertheless, to set a knowledge foundation, in this work, we focus on children as a broader user group.

inappropriate actions, repeating unsuitable words, or experiencing negative emotions evoked by certain songs [9, 61]. Today, children listen to music on digital devices via streaming platforms like Spotify or Apple Music [10, 67], of which a key component is their recommender system (**RS**): Streaming services provide next-song recommendations and prominently feature recommendations for playlists, albums, and songs on their main pages [82]. Hence, children are frequently presented with music curated by RS. Despite this, research on music RS for children often focuses on algorithmic performance or song preference [31, 81, 84], but rarely on what is *actually* befitting for children or what is a *good* item for them.

The pervasive nature of RS calls for action to grasp how their algorithms “construct taste and its implications to our cultural experiences and cultural processes, more broadly” [39]. This is essential for children, as not all the songs they like or encounter are meant for them [87]. For example, “Blurred Lines” by Robin Thicke, T.I., and Pharrell or “Despacito” by Luis Fonsi and Daddy Yankee are popular songs (900 million and 9 billion views on YouTube, respectively) with catchy rhythms children may find appealing [84]. Yet, they convey messages unsuitable for young audiences. RS may recognize the songs’ popularity and prioritize them in their suggestions, overlooking the diversity of audiences they reach. Considering these concerns, the need to further understand the interplay of RS and children is apparent. Doing so is of societal importance as children are in a forming stage with varying comprehension levels and skills and developing attitudes, behavior, and opinions [11, 12]. In this vulnerable phase [80], children require content that supports their development and does not harm them.

To set a foundation on the current understanding of how well RS cater to this demographic, we start by investigating the items RS might expose children to. Specifically, we focus on song **catalogs** that form the basis for recommendations [77] and the **lyrics** of these songs, which can convey messages not always befitting children. The former as a proxy for what is recommended, since regardless of the algorithms used, if the catalog includes items unsuitable for some audiences, the algorithm may reflect those in its recommendations. The latter, as prior work on limited samples revealed the presence in popular songs of explicit lyrics [35] (including profanity, vulgarity, and epithets [6]), misogyny [3, 4, 26], violence [46], and references to drugs and alcohol [47]. In this work, we address the RQ: ***Do song lyrics convey content befitting children?*** We empirically explore a large dataset from the Genius website², covering Pop, Rock, Rap, R&B, and Country songs, which we treat as a catalog that can be utilized for an RS. Given that children often engage with platforms not specifically designed for them [22, 49, 59]—where safeguarding mechanisms to limit inappropriate content may be absent or inconsistently applied—this catalog serves as a proxy for analyzing the types of items they may encounter through such systems. We conduct a content analysis on the lyrics using different **lenses** to assess whether items can be deemed befitting, accounting for multiple indicators that may ultimately

² <https://genius.com/>.

influence song suitability: curse words, profanity, offensive speech, hate speech, text complexity, and difficult-to-articulate words.

The main contribution of this work is a comprehensive overview of the lyrical content that RS may present to children, sparking reflections about how information access systems could handle content that may not be ideal for young audiences [87]. We publish our code in a public repository³.

Outcomes from this work have implications that extend beyond RS and music to various platforms, systems, and content types. As children’s access to information and media evolves, and mindful of safety-promoting practices, the question of whether presented content ‘fits’ becomes a must for the information retrieval (IR) community.

2 Background

Social learning theory suggests that children learn new behaviors by observing and imitating others [12] but also by adopting behaviors observed in media consumed [11]. This theory aligns with research on the influence of music on children, indicating that certain genres are associated with negative behavior of children [9,61]. Recent studies on smaller samples of music indicate that many popular songs from recent years contain explicit lyrics [35]. Aspects such as misogyny [3] and violence [46] are frequent occurrences in popular music. Themes such as drugs, alcohol, and sex [47] are most prevalent in Rap and R&B music.

The prominence of explicit lyrics and the potential negative effects of inappropriate lyrics raise concerns about what children are exposed to online. However, RS themselves tend not to filter inappropriate content well on prominent platforms. On the contrary, children are, in fact, likely to encounter inappropriate content on online platforms due to their recommender systems [69,70,94]. Algorithms that drive online platforms can even amplify potentially harmful content online in their attempts to maximize engagement with the platform [51,76,93]. Recent work highlights that recommendation algorithms not only reflect harmful characteristics present in the training corpora, such as stereotypes, but may even amplify their presence in recommendations delivered to children [88].

Despite these risks and initial explorations on how to counteract the spread of harmful recommendations [23], RS research on how children—a particularly vulnerable user group—can be affected by inappropriate recommendations is sparse. Instead, most RS research that has children as the protagonists focuses on uncovering children’s preferences [65,81,84] or improving the performance of RS [81,89]. Whether items are actually fitting to the children that use the systems is seldom explored [40,88], showcasing a critical gap. As a step toward considerations and safeguarding for children, thought has to be given to what it means for a recommendation to be truly fitting for a child.

³ https://github.com/JasperHeijne/appropriate_lyrics.git.

3 Experimental Setup

We treat the Genius lyrics database [55, 56]—GeniusDB—as a ‘catalog’ RS could use for recommendation purposes. GeniusDB is ideal for this study as it is sourced from the regularly updated Genius website and features a wide range of popular songs that children likely encounter when interacting with RS. It has 37,993 songs collected between September 2019 and January 2020, each annotated with its *lyrics* and at least one of five main *genres*: Pop, Rock, Country, Rap, and R&B. Table 1 shows the distribution and overlap of songs across genres (songs can be assigned multiple tags). The prevalence of Rap is anticipated, as Genius originated as Rap-Genius [52].

Table 1. Song distribution and genre overlap in GeniusDB.

Genre	Number of Songs	Overlap				
		<i>Rap</i>	<i>Pop</i>	<i>R&B</i>	<i>Rock</i>	<i>Country</i>
Rap	30,050	—	1,201	2,442	291	28
Pop	5,223	1,201	—	1,263	763	100
R&B	4,537	2,442	1,263	—	90	2
Rock	3,320	291	763	90	—	42
Country	164	28	100	2	42	—

To provide a comprehensive analysis of the words and expressions in each song lyric L in GeniusDB, we consider several lenses, each capturing a different facet that can be used to determine whether an item could be fitting for a child. We focus on identifying terms and phrases that could be considered harmful, particularly those that, if mimicked [11], may be insulting, offensive, or encourage negative attitudes toward others. We also account for *readability*-related considerations, which can be used to assess children’s understanding and the intended audiences of song lyrics.

Curse words are expletives, insults, blasphemy, obscenities, and racial slurs, which are considered inappropriate in social contexts. To identify curse words, we refer to Google’s list of banned curse words [37] that is updated monthly. As curse words indicators we use HIT_{Curse} , based on the HIT metric, which yields a 1 if L contains a curse word and 0 otherwise, and P_{Curse} , which captures the proportion of curse words in L . **Profanity** refers to hate speech, offensive language, and broader inappropriate language (e.g., toxic, obscene, threatening, and insulting) [27]. We use an SVM model [97] trained on 200k human-labeled samples of clean and profane strings to compute $L_{Profanity}$, a likelihood for L on a $[0, 1]$ range—the closer to 1, the more likely the song to be profane. This model, not based on a term list, enables a dynamic estimation leveraging context.

Offensive speech refers to rude or insulting expressions not targeting specific vulnerable subgroups. This differs from **hate speech** that “targets the ‘other’

in societies” [78], i.e., minority groups like racial, ethnic, religious, or cultural minorities, women, and the LGBTQI+ community. To detect these types of speech, we use two models [1, 2] that yield $L_{Offensive}$ and L_{HS} —likelihood scores on a $[0, 1]$ range, where 1 indicates a high likelihood of offensive or hate speech. These models, based on DeBERTa [45], were trained on the `tweet_eval` dataset [13] that includes tweets annotated with offensive language [95] and hate speech [14].

Readability refers to “a reader’s understanding, reading speed, and level of interest in the material” [25]. To infer the intended audience for L, we used widely used formulas. **Flesch-Kincaid** [34] assigns grade levels (ranging from grade 0 for Kindergarten to 13+ for graduate school) to texts based on shallow features like sentence length and syllables [5]. The **Dale-Chall** formula [21] considers familiarity with common words to assess readability. We also identify words with advanced orthographic and phonemic/phonetic patterns that we assume are **challenging** for a child to comprehend and, particularly, sing along to. For this, we use a model that accounts for phonemic decoding and educational development stages [73, 74]. With $P_{Challenging}$, we reflect the proportion of challenging words in L, excluding duplicate words and stopwords (as per NLTK’s English stopwords list⁴).

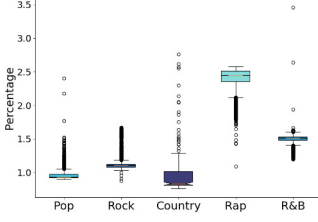
4 Results

We present the findings of our multi-lens analysis, with nuanced scrutiny of computed indicators per genre. We use a Kruskal-Wallis test ($p < .0001$) to compare indicators across genres and conduct post-Hoc Dunn’s test with Bonferroni correction ($p < .0001$) to show pairwise differences between genre pairs. Results from each lense per genre can be seen in Figs. 1 and 2.

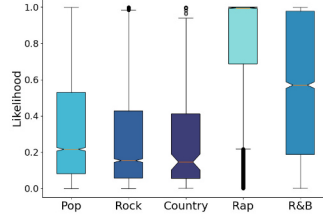
Curse words are frequent: HIT_{Curse} ranges from 23% for Country to 88% for Rap. P_{Curse} differs across genres (Fig. 1a); pairwise genre differences are significant, except for (Pop, Country) and (Rock, Country). Rap has the highest proportion of curse words—over twice as many as Pop, Rock, and Country. It is also the most likely genre to include **profanity** in its lyrics (Fig. 1b). Pairwise differences for $L_{Profanity}$ are significant for all genre pairs, except (Pop, Country) and (Rock, Country). **Offensive speech** trends are similar to those reported for curse words and profanity. Pairwise differences for $L_{Offensive}$ in Fig. 1c are significant, except for (Pop, Country) and (Rock, Country). As Fig. 1d highlights, results for **hate speech** differ from those previously reported. As per L_{HS} , Rock is the genre most likely to convey hate speech; pairwise differences between Rock and the other genres are significant. The only other significant differences are higher L_{HS} for Rap than Pop and R&B.

As per **Flesch-Kincaid** in Fig. 2a, Rap is the most difficult genre to comprehend. Other genres yield comparable readability levels; significantly higher

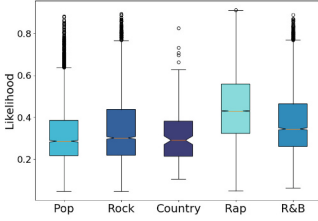
⁴ https://www.nltk.org/nltk_data/.



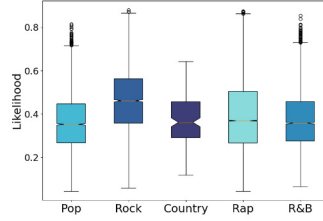
(a) P_{Curse} . Significant pairwise differences, except for (Pop, Country).



(b) $L_{Profanity}$. Pairwise differences are significant, except (Pop, Country) and (Rock, Country).



(c) $L_{Offensive}$. Pairwise differences are significant, except (Pop, Country) and (Rock, Country).

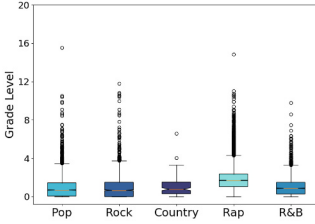


(d) L_{HS} . Significant pairwise differences between Rock and other genres, (Rap, Pop), and (Rap, R&B).

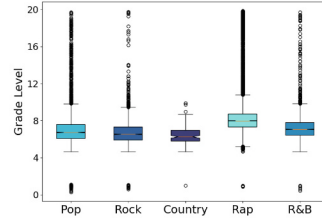
Fig. 1. Distributions of harms-related indicators for songs in GeniusDB, grouped by genre.

scores for R&B than Pop as an exception. **Dale-Chall** levels in Fig. 2b are significantly different across all genre pairs, except (Rock, Country). Regardless of genre, lyrics are relatively simple. For instance, the highest *Dale-Chall* level is ~ 8 which maps to 13–14 year olds. *Flesch-Kincaid* levels are even lower, ranging from ~ 0 to 2, i.e., 5–8 year olds would understand most lyrics according to this metric. $P_{Challenging}$ indicates that between 10 to 20% of the unique, non-stopwords in lyrics are **challenging** to utter and pronounce due to their advanced linguistic patterns [73,74]. In contrast to the readability scores, Rap is the genre with the least challenging words whereas Country songs include the most, followed by Rock.

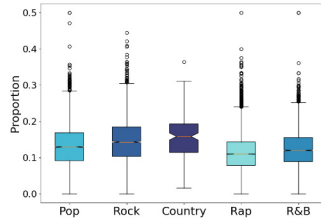
Overall, our findings point out that Rap emerges as the most ‘inappropriate’ genre given the prevalence of curse words, profanity, and offensive speech. The frequent use of curse words can lead children to mimic these terms and learn improper language [11]. Lyrics with a higher percentage of bad terminology, as seen for Rap music, could negatively influence their speech behavior. Although less prominent, we see similar trends for R&B, expected due to its large overlap with Rap (Table 1). Despite lower P_{Curse} , $L_{Profanity}$, and $L_{Offensive}$, Rock con-



(a) *Flesch-Kincaid*. Significant pairwise differences between Rap and all other genres and (R&B, Pop).



(b) *Dale-Chall*. Pairwise differences are significant, except (Country, Rock).



(c) $P_{Challenging}$. Pairwise differences are significant, except (Rock, Country).

Fig. 2. Distributions of readability-related indicators for songs in GeniusDB, grouped by genre.

veys hate speech, suggesting that this genre may contain a subtler form of undesirable content. Pop cannot be considered entirely child-friendly either; with over a third of Pop songs containing curse words, this genre requires careful screening to assess its suitability. Country, by contrast, shows few significant differences, which we attribute to a smaller sample size.

Consistently low readability levels evince that even young children could understand most lyrics in GeniusDB. Still, traditional readability formulas can be misleading [60], e.g., lyrics with short sentences lead to low levels due to the focus on shallow features like sentence length, but this does not imply these sentences are simple. Readability formulas fail to capture metaphors, euphemisms, or whether children can sing along to the songs they hear. $P_{Challenging}$ reveals a relatively low, but crucial number of challenging words in the lyrics, indicating that children can sing along to and utter many of the words in the lyrics. Lyrics from certain genres, particularly Rap, score the highest for most impropriety and reading difficulty indicators. Although more complex lyrics appear to lessen the risk of these genres influencing children, this is misleading. Despite the higher difficulty, the readability levels are quite low. This is concerning given the prevalence of inappropriate language and themes in these lyrics. The combina-

tion of accessible language and improper content may pose a risk. For instance, Rap songs, often seen as the hardest to understand, frequently use simple words children can easily repeat, unaware of the harmful content.

5 Discussion

Here we discuss emerging results (Sect. 4) and associated implications.

The Fit Concept. In addressing our RQ, we note that while songs are often deemed unsuitable by certain lenses, determining *truly* befitting songs is challenging. The ‘fit’ of songs for children hinges on their content. Those promoting violence, stereotypes, or hate speech may directly impact behavior [48, 75, 93]. Even without explicit offensive language, some songs may subtly reinforce biases against ethnicities, religions, or social groups [12, 15]. Children, however, are not a homogeneous group—they vary in developmental stages [32]. Their susceptibility to content depends on age, gender, and previous experiences [7]. Hence, a child’s individual characteristics also determine the ‘fit’ of an item. Further, external factors play a role. For instance, “Blurred Lines” might promote problematic views on consent [11]. It can also be a vehicle for parent-child discussions on the topic [16], as discussions with parents or caretakers can mitigate potential harm [66]. Further, what may appear as hate speech, at first sight, could in reality be reclaimed language—derogatory terms, which instead of being harmful, may be used to refer positively to in-groups for de-stigmatization of negative group labels [28, 38]. Appropriateness is not a stable construct: for instance, it may vary across different cultural contexts; something that is acceptable in one culture may not be in another. An item that is unsuitable for a child of a certain age may become suitable for them when they get older and gain new experiences. For example, the sexual connotation in songs like “Despacito” may be considered inappropriate for younger children or in certain cultures, but may become less problematic as children grow older and become more familiar with such themes.

Ultimately, defining ‘fit’ for children is complex; it requires simultaneously considering the inherent song traits, the characteristics of the listener, and the listening context. Lyrics are a relevant starting point for deepening the understanding of music characteristics [71]; however, these three pillars—*item*, *user*, and *context*—are also applicable to domains beyond music [90]. Thorough assessment methods of *what* is appropriate for *whom* require the involvement of experts who understand child development, cultural norms, and media effects.

Practical Considerations for RS. Our analysis reveals that many songs in GeniusDB contain words and phrases that could be perceived as improper. An RS leveraging this catalog may naturally suggest such songs. This issue is not limited to music; similar patterns have been noted in recommendation models across various domains [70, 93]. Unfortunately, ‘fit’ is rarely part of *evaluation* frameworks, which typically focus on accuracy metrics [42]. Although modern evaluations increasingly account for diversity, serendipity, and fairness [30, 42],

assessing whether an item genuinely suits a user group is often overlooked [33]. Multistakeholder evaluation research [19, 40] echoes the need to look beyond accuracy to consider other needs and requirements for assessment purposes. The lenses we considered offer a foundation for evaluation criteria beyond mere accuracy—an especially critical step for safeguarding vulnerable children as they navigate online content. With online platforms playing an increasingly influential role in shaping how young users view the world and assess right from wrong, unsuitable or inappropriate content can directly impact their development [68]. Our work emphasizes the need for safeguarding methods, parental controls, and content filtering mechanisms to consider multiple lenses and nuanced consideration of appropriateness.

To contribute to healthy media consumption, RS could remove items deemed ‘unsafe,’ similar to platforms using “parental advisory” labels to identify and remove explicit content [50, 85]; or leverage editorial curation to offer “child-friendly” songs [36, 86]. These types of *filtering*, however, are known to fall short [8], overlooking the subtleties of what makes a song unsuitable for a child in a given context. Binary classifications for fit are overly simplistic in practice. Simply ‘hiding’ potentially inappropriate content may not protect children [58]. Fostering media literacy [18] and a human-driven approach to design offers a better alternative. RS can optimize item selections to fulfill various roles, from providing emotional support and relaxation to entertaining [17]. Additionally, these tailored RS can facilitate positive interactions with potentially challenging content [66]. *Designing* recommendation architectures that explicitly incorporate as part of the recommendation-generation process indicators like the ones we considered to model items, alongside the child’s characteristics, demographics, and contextual factors, could result in systems that promote not only enjoyable but also appropriate and beneficial content for children’s specific situations.

In *real-world scenarios*, RS reach diverse audiences, including minority groups who are inevitably exposed to suggestions on these platforms. Research shows that algorithms can negatively impact minority groups, resulting in unfair outcomes for them, enabled through biases and discrimination [63]. System design should address factors relevant to minorities like personality, emotional state, or situation [83] and account for how RS decisions affect minorities. A one-size-fits-all approach will not solve this problem. Instead, designing RS with vulnerable users in mind will prioritize user well-being [29] and foster a benevolent and inclusive approach.

Considerations for IR. Our findings extend beyond RS to broader IR tasks, as every information access system inherently involves *content* presentation. Search engines, for example, have addressed issues like the usability and understandability of results for children, and mitigating ranking biases related to protected attributes such as gender [64, 96]. Still, the lenses we introduce offer additional dimensions to scrutinize potential harms that may affect the experience of ‘non-conventional’ user groups [72]. Media—from informational to entertaining—is consumed in increasingly dynamic and personalized ways across search systems, RS, or conversational interfaces [24, 53]. Moreover, automatically generated con-

tent is becoming more prevalent [20,98] and generative models are integrated within retrieval systems [79]. These increasingly complex dynamics exacerbate uncertainties regarding *what* content is presented to users. Information access and retrieval systems accessed by a wide range of users must be aware of the potential harms their methods may promote. Content analysis plays a key role in critically probing item corpora and informing the design of mechanisms that enable safer interactions. In our work, we leverage natural language processing methods on text as a starting point; more holistic probing must account for a wider range of media, such as images and audio, as well as gauging algorithms that curate and present content to ensure they do not expose users to what may not be suitable for them.

Limitations and Future Work. In GeniusDB, there is a notable imbalance in genre representation, with Rap being the most frequent genre. While we assume that observed trends for prominent genres may reflect broader patterns within those genres, this may not be true for underrepresented ones. For instance, the genre ‘Country’ is represented by only 164 songs. To improve the robustness and ensure the generalizability of our findings, future work should consider incorporating additional data sources with more balanced genre distributions.

Further, some lyrics mix foreign languages with English; others are mostly in foreign languages with scattered English words. We focused on English and selected text analysis models accordingly. We plan to conduct multilingual scrutiny in the future. While models from natural language processing are an effective method to detect inappropriate content, they may be disadvantaged for low-resource languages [54], emphasizing the need for complementary approaches that do not rely solely on language-specific resources. These can include metadata-based signals or analysis of visual content that often accompanies musical items.

As many factors shape lyric analysis, future work should broaden the scope to other types of inappropriate content beyond the ones we considered. This, however, must be approached with care. The perceived appropriateness of themes like violence or misogyny, is not easily mapped to specific age groups and may vary across cultural, social, and individual boundaries [62], highlighting the need for nuanced and reflexive frameworks when assessing suitability for children. Lenses should also address content not easily inferred from naive word analysis, such as metaphors or stereotypes, which may subtly convey harmful messages despite not containing explicitly inappropriate language.

Our study indicates that harmful lyrics *can* be promoted by RS when based on popular item corpora, as a large majority of lyrics include content that can be unfitting to some users. However, the degree to which RS expose children to potentially unfitting items remains unexplored. Future work should study children’s exposure to such items due to RS [cf. 88] to analyze the extent to which RS can negatively impact children.

6 Concluding Remarks

We have provided an analysis of the content prevalent in popular music genres. Our results indicate that inappropriate content exists across all five music genres and that children can understand and articulate most lyrics. This, in turn, suggests that not all contemporary music that RS are likely to recommend may not be suitable for children, highlighting the need for caution in RS and other information access systems targeting young users, particularly considering the nuanced aspects influencing item ‘fit’. With this work, we aim to raise awareness about the challenges that information access systems face when curating suggestions, particularly in contexts involving children—whether as a primary audience or indirectly impacted users [19]. We encourage reflections on the design, evaluation, and deployment of these systems. Fostering transparency and attentiveness to the content provided can contribute to a more responsible and inclusive digital music environment and beyond, ultimately supporting safer online platforms for all users.

References

1. KoalaAI/HateSpeechDetector · Hugging Face — huggingface.co. <https://huggingface.co/KoalaAI/HateSpeechDetector> (2023), Accessed 12 09 2024
2. KoalaAI/OffensiveSpeechDetector · Hugging Face — huggingface.co. <https://huggingface.co/KoalaAI/OffensiveSpeechDetector> (3), Accessed 12 09 2024
3. Adams, T., Fuller, D.: The words have changed but the ideology remains the same: misogynistic lyrics in rap music. *JBS* **36**(6), 938–957 (2006)
4. Adams, T.M., Fuller, D.B.: The words have changed but the ideology remains the same: misogynistic lyrics in rap music. *J. Black Stud.* **36**(6), 938–957 (2006)
5. Allen, G., Milton, A., Wright, K.L., Fails, J.A., Kennington, C., Pera, M.S.: Supercalifragilisticexpialidocious: why using the “right” readability formula in children’s web search matters. In: *European Conference on Information Retrieval*, pp. 3–18. Springer (2022)
6. Amdan, N.F.B., Shaari, A.H.: An analysis of profanity in English lyrics. *J. Wacana Sarjana* **1**(1) (2017)
7. Anderson, C., et al.: The influence of media violence on youth. *PSPI* **4**(3), 81–110 (2003)
8. Anuyah, O., Milton, A., Green, M., Pera, M.S.: An empirical analysis of search engines’ response to web search queries associated with the classroom setting. *Aslib* **72**(1), 88–111 (2020)
9. Arnett, J.J.: Heavy metal music and reckless behavior among adolescents. *J. Youth Adolesc.* **20**(6), 573–592 (1991)
10. Auxier, B., Anderson, M., Perrin, A., Turner, E.: Children’s engagement with digital devices, screen time. <https://www.pewresearch.org/internet/2020/07/28/childrens-engagement-with-digital-devices-screen-time/> (2021), Accessed 23 Jul 24
11. Bandura, A.: Social cognitive theory of mass communication. In: *Media effects*, pp. 110–140. Routledge (2009)
12. Bandura, A., Walters, R.H.: *Social learning theory*, vol. 1. Englewood cliffs Prentice Hall (1977)

13. Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., Neves, L.: TweetEval: unified benchmark and comparative evaluation for tweet classification. In: Findings of EMNLP (2020)
14. Basile, V., et al.: SemEval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter. In: SemEval, pp. 54–63. ACL (2019)
15. Block, K., Gonzalez, A.M., Choi, C.J., Wong, Z.C., Schmader, T., Baron, A.S.: Exposure to stereotype-relevant stories shapes children’s implicit gender stereotypes. *PLoS ONE* **17**(8), e0271396 (2022)
16. Blossom, P.: 15 moms share the dirty, inappropriate song their kid is obsessed with. <https://www.romper.com/p/15-moms-share-the-dirty-inappropriate-song-their-kid-is-obsessed-with-5467> (2016), Accessed 25 Jul 24
17. Boal-Palheiros, G.M., Hargreaves, D.J.: Listening to music at home and at school. *Br. J. Music Educ.* **18**(2), 103–118 (2001)
18. Buckingham, D.: Digital media literacies: rethinking media education in the age of the internet. *Res. Comp. Int. Educ.* **2**(1), 43–55 (2007)
19. Burke, R., et al.: De-centering the (traditional) user: multistakeholder evaluation of recommender systems. arXiv preprint [arXiv:2501.05170](https://arxiv.org/abs/2501.05170) (2025)
20. Cao, Y., et al.: A comprehensive survey of ai-generated content (aigc): a history of generative ai from gan to chatgpt. arXiv preprint [arXiv:2303.04226](https://arxiv.org/abs/2303.04226) (2023)
21. Chall, J.S., Dale, E.: Readability Revisited: The New Dale-Chall Readability Formula. Brookline Books (1995)
22. Charmaraman, L., Lynch, A.D., Richer, A.M., Grossman, J.M.: Associations of early social media initiation on digital behaviors and the moderating role of limiting use. *Comput. Hum. Behav.* **127**, 107053 (2022)
23. Chee, J., Kalyanaraman, S., Ernala, S.K., Weinsberg, U., Dean, S., Ioannidis, S.: Harm mitigation in recommender systems under user preference dynamics. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 255–265 (2024)
24. Chen, J., et al.: When large language models meet personalization: perspectives of challenges and opportunities. *World Wide Web* **27**(4), 42 (2024)
25. Collins-Thompson, K.: Computational assessment of text readability: a survey of current and future research. *ITL* **165**(2), 97–135 (2014)
26. Couto, L., Hust, S.J., Rodgers, K.B., Kang, S., Li, J.: A content analysis of music lyrics exploring the co-occurrence of violence, sexual content, and degrading terms toward women. *Sex. Cult.* **26**(6), 1965–1980 (2022)
27. Davidson, T., Warmesley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: AAAI ICWSM, pp. 512–515 (2017)
28. De Smet, B., Dhaenens, F.: Heteromascularity and queer reappropriation in music streaming practices: exploring homonegative curation by ordinary spotify users. *Fem. Media Stud.* **24**(4), 695–712 (2024)
29. Ekstrand, M., Beattie, L., Pera, M.S., Cramer, H.: Not just algorithms: strategically addressing consumer impacts in information retrieval. In: ECIR, pp. 314–335. Springer (2024)
30. Ekstrand, M., Das, A., Burke, R., Diaz, F.: Fairness in recommender systems. In: Recommender Systems Handbook, pp. 679–707. Springer (2012)
31. Ekstrand, M., et al.: All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In: ACM FAccT Conference, pp. 172–186. PMLR (2018)
32. Feldman, D.: Piaget’s stages: the unfinished symphony of cognitive development. *New Ideas Psychol.* **22**(3), 175–231 (2004)

33. Ferraro, A., Ferreira, G., Diaz, F., Born, G.: Measuring commonality in recommendation of cultural content: recommender systems to enhance cultural citizenship. In: ACM RecSys, pp. 567–572 (2022)
34. Flesch, R.: A new readability yardstick. *J. Appl. Psychol.* **32**(3), 221 (1948)
35. Frisby, C.M., Behm-Morawitz, E.: Undressing the words: prevalence of profanity, misogyny, violence, and gender role references in popular music from 2006–2016. *Media Watch* **10**(1) (01 2019)
36. Gabb Wireless Inc.: Gabb Music — gabb.com. <https://gabb.com/gabb-music> (2024), Accessed 12 Sep 2024
37. Gabriel, R.J.: Full list of bad words and top swear words banned by Google. <https://github.com/coffee-and-fun/google-profanity-words> (2024), Accessed 23 May 24
38. Galinsky, A.D., Wang, C.S., Whitson, J.A., Anicich, E.M., Hugenberg, K., Bodenhausen, G.V.: The reappropriation of stigmatizing labels: the reciprocal relationship between power and self-labeling. *Psychol. Sci.* **24**(10), 2020–2029 (2013)
39. Gaw, F.: Algorithmic logics and the construction of cultural taste of the netflix recommender system. *Media Cult. Soc.* **44**(4), 706–725 (2022)
40. Gómez, E., Charisi, V., Chaudron, S.: Evaluating recommender systems with and for children: towards a multi-perspective framework. In: *Perspectives@ RecSys* (2021)
41. Gregory, A.H., Worrall, L., Sarge, A.: The development of emotional responses to music in young children. *Motiv. Emot.* **20**, 341–348 (1996)
42. Gunawardana, A., Shani, G., Yogev, S.: Evaluating recommender systems. In: *Recommender Systems Handbook*, pp. 547–601. Springer (2012)
43. Hallam, S.: The power of music: its impact on the intellectual, social and personal development of children and young people. *IJME* **28**(3), 269–289 (2010)
44. Hargreaves, D.J., North, A.C., Tarrant, M.: How and why do musical preferences change in childhood and adolescence. *The child as musician: A handbook of musical development*, pp. 303–322 (2015)
45. He, P., Liu, X., Gao, J., Chen, W.: Deberta: decoding-enhanced bert with disentangled attention. [arXiv:2006.03654](https://arxiv.org/abs/2006.03654) (2020)
46. Herd, D.: Changing images of violence in rap music lyrics: 1979–1997. *J. Public Health Policy* **30**(4), 395–406 (2009)
47. Holody, K., Anderson, C., Craig, C., Flynn, M.: “Drunk in love”: the portrayal of risk behavior in music lyrics. *J. Health Commun.* **21**(10), 1098–1106 (2016)
48. Huesmann, L.R., Moise-Titus, J., Podolski, C.L., Eron, L.: Longitudinal relations between children’s exposure to tv violence and their aggressive and violent behavior in young adulthood: 1977–1992. *Dev. Psychol.* **39**(2), 201–221 (2003)
49. Huk, T., et al.: Use of facebook by children aged 10–12. presence in social media despite the prohibition. *New Educ. Rev.* **46**(1), 17–28 (2016)
50. Jackson, A.: Parental advisory: the story of a warning label (09 2020), <https://daily.jstor.org/parental-advisory-the-story-of-a-warning-label/>
51. Kasirzadeh, A., Evans, C.: User tampering in reinforcement learning recommender systems. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 58–69 (2023)
52. Kehrer, L.: Genius (formerly rap genius). genius media group, inc. genius. com. *J. Soc. Am. Music* **10**(4), 518–520 (2016)
53. Klopfenstein, L.C., Delpriori, S., Malatini, S., Bogliolo, A.: The rise of bots: a survey of conversational interfaces, patterns, and paradigms. In: *Proceedings of the 2017 Conference on Designing Interactive Systems*, pp. 555–565 (2017)

54. Li, Z., Shi, Y., Liu, Z., Yang, F., Payani, A., Liu, N., Du, M.: Language ranker: a metric for quantifying llm performance across high and low-resource languages. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 28186–28194 (2025)
55. Lim, D., Benson, A.: Expertise and dynamics within crowdsourced musical knowledge curation: a case study of the genius platform. In: AAAI ICWSM, vol. 15, pp. 373–384 (2021)
56. Lim, D., Benson, A.R.: Genius - expertise dataset — cs.cornell.edu. <https://www.cs.cornell.edu/~arb/data/genius-expertise/> (2021), Accessed 12 Sep 2024
57. Lincoln, S.: Feeling the noise: teenagers, bedrooms and music. *Leis. Stud.* **24**(4), 399–414 (2005)
58. Livingstone, S., Helsper, E.J.: Parental mediation of children’s internet use. *J. Broadcast. Electr. Media* **52**(4), 581–599 (2008)
59. Livingstone, S., Ólafsson, K., Staksrud, E.: Social networking, age and privacy (2011)
60. Madrazo Azpiazu, I., Pera, M.S.: Multiattentive recurrent neural network architecture for multilingual readability assessment. *TACL* **7**, 421–436 (2019)
61. Martin, G., Clarke, M., Pearce, C.: Adolescent suicide: music preference as an indicator of vulnerability. *J. Am. Acad. Child Adolesc. Psychiatry* **32**(3), 530–535 (1993)
62. McClelland, S.I., Hunter, L.: Bodies that are always out of line: a closer look at “age appropriate sexuality”. In: *The Moral Panics of Sexuality*, pp. 59–76. Springer (2013)
63. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM CSUR* **54**(6), 1–35 (2021)
64. Milton, A., Anuya, O., Spear, L., Wright, K.L., Pera, M.S.: A ranking strategy to promote resources supporting the classroom environment. In: 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 121–128. IEEE (2020)
65. Milton, A., Batista, L., Allen, G., Gao, S., Ng, Y.K.D., Pera, M.S.: “Don’t judge a book by its cover”: Exploring book traits children favor. In: Proceedings of the 14th ACM Conference on Recommender Systems, pp. 669–674 (2020)
66. Nathanson, A.: Identifying and explaining the relationship between parental mediation and children’s aggression. *CR* **26**(2), 124–143 (1999)
67. OfCom, U.: Children and Parents: Media use and Attitudes Report 2022. Office of Communications London, London (2022)
68. O’Keeffe, G.S., Clarke-Pearson, K., et al.: The impact of social media on children, adolescents, and families. *Pediatrics* **127**(4), 800–804 (2011)
69. Papadamou, K., et al.: Disturbed youtube for kids: characterizing and detecting inappropriate videos targeting young children. *AAAI ICWSM* **14**(1), 522–533 (2020)
70. Papadamou, K., et al.: Disturbed youtube for kids: characterizing and detecting inappropriate videos targeting young children. In: AAAI ICWSM, vol. 14, pp. 522–533 (2020)
71. Parada-Cabaleiro, E., et al.: Song lyrics have become simpler and more repetitive over the last five decades. *Sci. Rep.* **14**(1), 5531 (2024)
72. Pera, M.S., Cena, F., Huibers, T., Landoni, M., Mauro, N., Murgia, E.: 1st workshop on information retrieval for understudied users (ir4u2). In: European Conference on Information Retrieval, pp. 409–414. Springer (2024)

73. Pinney, C., Bettencourt, B.J., Fails, J.A., Kennington, C., Wright, K.L., Pera, M.S.: How readability cues affect children's navigation of search engine result pages. In: ACM IDC, pp. 62–69 (2024)
74. Pinney, C., Kennington, C., Pera, M.S., Wright, K.L., Fails, J.A.: Incorporating word-level phonemic decoding into readability assessment. In: LREC-COLING, pp. 8998–9009 (2024)
75. Raj, A., Milton, A., Ekstrand, M.D.: Pink for princesses, blue for superheroes: The need to examine gender stereotypes in kid's products in search and recommendations. In: 5th International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems (KidRec 2021), in Conjunction with ACM IDC 2021, <https://doi.org/10.48550/arXiv.2105.09296> (2021)
76. Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A., Meira Jr, W.: Auditing radicalization pathways on youtube. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 131–141 (2020)
77. Ricci, F., Rokach, L., Shapira, B.: Recommender systems: techniques, applications, and challenges. In: Recommender Systems Handbook, pp. 1–35 (2021)
78. Rights for Peace: what is hate speech? <https://www.rightsforpeace.org/hate-speech#:~:text=Speech%20that%20is%20simply%20offensive> (2024), Accessed 23 Jul 24
79. Salemi, A., Zamani, H.: Evaluating retrieval quality in retrieval-augmented generation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2395–2400 (2024)
80. Scharfe, E.: Development of emotional expression, understanding, and regulation in infants and young children. In: The Handbook of Emotional Intelligence: Theory, Development, Assessment, and Application at Home, School, and in the Workplace, pp. 244–262 (2000)
81. Schedl, M., Bauer, C.: Online music listening culture of kids and adolescents: Listening analysis and music recommendation tailored to the young. In: KidRec Workshop (2017). <https://doi.org/10.48550/arXiv.1912.11564>
82. Schedl, M., Knees, P., McFee, B., Bogdanov, D.: Music recommendation systems: techniques, use cases, and challenges. In: Recommender Systems Handbook, pp. 927–971. Springer (2021)
83. Schedl, M., Zamani, H., Chen, C.W., Deldjoo, Y., Elahi, M.: Current challenges and visions in music recommender systems research. *Int. J. Multimedia Inf. Retrieval* **7**, 95–116 (2018)
84. Spear, L., et al.: Baby shark to barracuda: analyzing children's music listening behavior. In: ACM RecSys, pp. 639–644 (2021)
85. Spotify AB: explicit content - Spotify — support.spotify.com. <https://support.spotify.com/us/article/explicit-content/> (2024), Accessed 12 Sep 2024
86. Spotify AB: spotify kids - spotify (US) — spotify.com. <https://www.spotify.com/us/kids/> (2024), Accessed 12 Sep 2024
87. Tang, T.Y., Winoto, P.: I should not recommend it to you even if you will like it: the ethics of recommender systems. *NRHM* **22**(1–2), 111–138 (2016)
88. Ungruh, R., Al Nahadi, M., Pera, M.S.: Mirror, mirror: exploring stereotype presence among top-n recommendations that may reach children. *ACM Trans. Recommender Syst.* (2025)
89. Ungruh, R., Bellogín, A., Pera, M.S.: The impact of mainstream-driven algorithms on recommendations for children. In: European Conference on Information Retrieval, pp. 67–84. Springer (2025)

90. Ungruh, R., Pera, M.S.: Ah, that's the great puzzle: on the quest of a holistic understanding of the harms of recommender systems on children. In: DCDW co-located with ACM IDC 2024. <https://doi.org/10.48550/arXiv.2405.02050> (2024)
91. UNICEF: The convention on the rights of the child: the children's version (2019), <https://www.unicef.org/child-rights-convention/convention-text-childrens-version#:~:text=A%20child%20is%20any%20person%20under%20the%20age%20of%2018>
92. Valkenburg, P.M., Cantor, J.: The development of a child into a consumer. *J. Appl. Dev. Psychol.* **22**(1), 61–72 (2001)
93. Whittaker, J., Looney, S., Reed, A., Votta, F.: Recommender systems and the amplification of extremist content. *Internet Policy Rev.* **10**(2) (2021)
94. Yesilada, M., Lewandowsky, S.: Systematic review: Youtube recommendations and problematic content. *Internet Policy Rev.* **11**(1), 1652 (2022)
95. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Semeval-2019 task 6: identifying and categorizing offensive language in social media (offenseval). In: *SemEval*, pp. 75–86 (2019)
96. Zerveas, G., Rekabsaz, N., Cohen, D., Eickhoff, C.: Mitigating bias in search results through contextual document reranking and neutrality regularization. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2532–2538 (2022)
97. Zhou, V.: Building a better profanity detection library with scikit-learn, February 2019, <https://victorzhou.com/blog/better-profanity-detection-with-scikit-learn/>
98. Zhu, Y., et al.: Large language models for information retrieval: a survey. *arXiv preprint arXiv:2308.07107* (2023)